FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Muhammad Waqas

FEATURE GENERATION FOR OPTIMIZATION OF MARKETING CAMPAIGNS

Master's Thesis
Degree Programme in Computer Science and Engineering
June 2020

Waqas M. (2020) Feature Generation for Optimization of Marketing Campaigns.
University of Oulu, Degree Programme in Computer Science and Engineering, 40 p.

ABSTRACT

Utilizing the gaming data for optimizing the entire gaming paradigm has revolutionized the thought process of developers and gamers alike. The significance of the gaming data can be judged from the fact that it is being used productively by the marketing agencies to develop algorithms that could predict the behavior of a certain gamer and the reaction to updates. The core idea behind the solution proposed and implemented in this thesis is focused on making the marketing campaigns more impactful. According to the facts from credible online resources, i.e., Statista.com, the business-to-business (B2B) organizations spent over $12.3 billion on marketing campaigns. Since one of the major aims of a marketing campaign is customer acquisition, which is also referred to as demand generation, measuring the success rate of the marketing campaign is also of great importance. Besides, the conventional Customer Relation Managers (CRMs) don't have such features using which, the businesses can monitor the effectiveness of the marketing campaigns. The system this thesis proposes aims to analyze the gaming data, which can be used to extract features for refined marketing campaigns. To analyze and precisely classify the gaming data, this thesis proposes an algorithm running behind a full-fledged marketing campaign that can yield optimal results and which can be further refined to predict the future purchase behavior of the users in such marketing campaigns. To accomplish this task, the Random Forest Classifier is the one, which this thesis proposes and has been implemented to optimize feature selection in order to enhance the profit revenue of the business. The promising results of empirical research and studies have proven the capability of the random forest classifier, and after employing it in the research, it has been established that the mentioned classifier is absolutely capable of extracting significant features on the basis of the gaming data sets that were provided. More importantly, this study has indicated that the Random Forest classifier gives better results in predicting the purchase likelihood, which is an essential milestone for our project. It should be noted that the solution we have proposed does not only serve to predict the purchase likelihood, but it can also be preferably utilized for other aims and objectives which are related to optimizing the marketing campaigns.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS

k	Number of Features
N	Number of Samples
CSE	Computer Science and Engineering
MQLs	Marketing Qualified Leads
SQLs	Sales Qualified Leads
USPs	Unique Selling Points
RFM	recency,frequency monetary
ML	Machine Learning
XGBoost	Extreme Gradient Boosting
OLAP	online analytical processing
DNA	Deoxyribonucleic acid
RFID	Radio-frequency identification
Wi-Fi	wireless fidelity
MLE	Maximum likelihood estimation
LR	Linear Regression
RF	Random Forest
SVM	Support vector Machine
SVR	Support vector Regression
DT	Decision Tree
SL	Supervised Learning
FS	Feature Selection
HH	Head Head
TT	Tail Tail
TH	Tail Head
HT	Head Tail
FSH	Forward Selection Heuristics
BSH	Backward Selection Heuristics
SMEs	Small and Medium Enterprises
CFS	Correlation Feature Selection
SCFC	Subset Consistency Feature Selection
BQ	Big Query
TN	True positive
FP	False Positive
FN	False Negative
TN	True negative
DFS	Deep Feature Synthesis
AI	Artificial Intelligence
PMF	Probability Mass Function
K-NN	k-nearest neighbors
ANOVA	Analysis of variance
2D	Two dimensional
RFE	Recursive Feature Elimination
FDR	False directory rate
CER	Conditional error rate

CART	Classification and Regression Trees
ROC	Receiver operating characteristic
CV	Cross Validation
CRMs	Customer Relation Managers
SOEP	Social Organization for Environmental Protection Copyright

FOREWORD

It has been a long journey with various highs and lows. It was not always a smooth ride starting from the beginning and coming to an end. After years of research, hard work, and dedication, I have been able to complete my thesis. This feat was not accomplished solely by me, for I was fortunate and privileged to have worked with some wonderful people who were responsible for pushing and motivating me to complete my thesis.

Working on my thesis had been challenging as I navigated a sea of research papers and technical documents necessary in completing it. The world is ever-evolving, and that had a significant impact on the direction I took in writing my thesis. I am a person that is driven towards finding a solution for the betterment of humanity. I am passionate about the advancement of artificial intelligence and its utilization in making life easier for the world. This thesis is evidence of all the skills and knowledge that I have acquired, and it also breaks down my plans to use these ideas and theories to make a global impact.

Of course, I did not journey through this scholastic endeavor all by myself. I had reputable and dignified mentors who were paramount in the completion of this thesis.

First, I would like to acknowledge Mr. Seppo Kangas, who served as my technical supervisor and helped me enormously in generating the necessary datasets that I required in my research. He always helped in perusing my technical work and was still handy to give me the appropriate corrections that were necessary for improving my work.

Next, I have also to acknowledge Mrs. Susanna Pirttikangas, my second supervisor. This is a woman with a very incredible and fascinating brain. She was incredibly supportive of my work and was useful in providing the much-needed insight for my work. Her excellent remarks and feedback's helped me improve on my work as I compiled hours upon hours of scientific materials for the thesis.

Mrs. Susanna assisted me in perfecting the structure of my thesis. She never relented in answering the phone or indulging in a video call in order to provide me with the clarity that I needed at that particular point in time. His willingness to supervise without issue was a huge bonus to me.

Lastly, but not least, I would like to recognize Mrs. Ella Peltonen for her unrelenting tenacity in supporting me with carrying out my research for my thesis. She was always there every step of the way, fact-checking and proofreading every new piece of information with independent scrutiny.

All these people were tantamount to the execution and completion of this project, and I would have probably been unable to finish this without their invaluable input. I want to appreciate their help and use all that I have learned to improve the world around me.

Oulu, June 16th, 2020

Muhammad Waqas

1. INTRODUCTION

Marketing campaigns are a way to promote brands by leveraging the power of digital media, for example, television, social media platforms, and print media. The advent of technology has not only brought about a significant change in the way people do business. It has also revamped the marketing paradigm as a whole. The success and failure of every marketing campaign are not only reliant on the budget anymore since various factors play a pivotal role in converting visitors of online services to potential customers. Measuring the success rate has now become precise and real-time due to the inclusion of the latest technological flavors in the field of marketing, machine learning, artificial intelligence, and data mining, to be exact.

Owing to the efficiency of the fields mentioned above, it has become easier to identify marketing qualified leads (MQLs) and sales qualified leads (SQLs), which further aid in optimizing the marketing campaigns eventually leading to profitable returns. Since every business has a particular target audience having a list of unique selling points (USPs), the marketing teams have to perform a lot of research whenever a new product comes their way. Having a machine learning-powered real-time data analysis system running on-board, the time which is spent on R&D, can be minimized and can be used productively elsewhere. With this idea in mind, defining the sales projections of specific customer segments via RFM (recency, frequency and monetary) modeling within applications and campaigns which are running on ML-based systems, is becoming pervasive.

This research, specifically, has been conducted primarily using the data from a game having more than 500 million downloads and an exceptional rating of $4.4/5$ at the play store. For data this big, there can be hundreds and thousands of user stories which the marketers can target in order to make their upcoming campaign more impactful. The data-set gathered for this study comprises millions of instances. One of the main aims of the study is to extract impactful features that could further assist in predicting the purchase behavior of digital product consumers. For such a voluminous amount of data, the data needs to be pre-processed using different techniques. This is an exhaustive task since it requires constant monitoring.

After completing the preprocessing step features are generated with careful analysis. Feature generation is a method of generating distinct features from one or multiple, already existing features. This process adds new information to be accessible during the model construction and, therefore, can result in a comparatively accurate model. Right after features generation, this study employs a data classification technique that predicts the purchase behavior of the user.

This research is data-driven, and the techniques employed are solely to extract distinct features that could lead to the completion of the main objective which is, to develop a system that could predict a gamer's purchase behavior while playing the game. For a successful marketing campaign that aims to target a certain type of user base in online gaming the prediction is next to impossible unless the marketer who has been tasked with designing a specific campaign is well-aware of the behavior of that class of users. Basically, this study has catered to the needs of the domain of digital marketing using contemporary data classification models. The solution that has been proposed as the outcome of this study is a predictive model that is capable of

a predicting the user's purchase behavior owing to the features which were extracted using refined dataset.

1.1. Scope of Thesis

This study has aimed to equip a marketing campaign with such a capability that it can predict the purchase behavior of a digital product consumer. The problem which became the food for thought for this work was derived from the fact that the product owners were finding it quite challenging to figure out the performance of marketing campaigns. Even more so, predicting the success factor of a marketing campaign to be run in the future was also becoming increasingly uncertain. The automated marketing platforms are quite capable of monitoring marketing campaigns, but they are not yet equipped with such algorithms, which can help in monitoring multiple marketing campaigns. Besides, such marketing campaigns do not give any information about consumer behavior. To ease out the scenario of successfully predicting the purchase behavior of consumers by running impactful marketing campaigns, this work explains the development of a predictive model built with random forest classifier after extracting distinct features.

The features for the study have been identified by successfully classifying the data along with several features and then selecting capable ordered pairs. The process was deliberately repeated in order to derive such features that possess the ability to predict the purchase behavior of digital product consumers. There were two aims of the study. First, to link the users by training the model on their set of behaviors second, to train the model in such a way that it can successfully predict which marketing campaign would work effectively with what type of users. The traditional approach would be to run several campaigns to figure out the answers to these questions, but it would have cost a lot of time and budget. Even more so, the marketers would still be in confusion, in fact, an anomaly could happen at any time along the way.

Following are the research questions which have been addressed as part of this study:

1. Whether or not a player would purchase the premium features offered in a game?
2. What features would play a crucial role in making the purchase decision?
3. Could these features be used to develop a predictive model?

The answers to the question raised above, could further lead to developing optimized digital marketing campaigns without having to spend extra amounts on trial campaigns thus saving time and the budget.

2. LITERATURE REVIEW

Limiting the chances of uncertainty and increasing the probability of successful prediction is one of the significant aims of data sciences in general, machine learning, and data mining to be specific [1]. For the researchers working in such fields, the primary source of training, a classifier is the data. When it comes to the digital marketing paradigm, the data can be in many forms. Machine learning and data mining are such fields that have brought about a change in the world of marketing. The uncertainties involved which are related to predicting users purchase behavior are being lessened owing to the on-going research.

Figure 1. General pipeline of the feature selection

Due to its significance in many daily life applications, we resort to data mining and machine learning to extract useful patterns that could help in predicting the purchase behavior of the customers in online gaming. In Figure 1, we show a general pipeline of the feature selection process that we use in our work. Consequently, the following sections are organized in a way to discuss the relevant technologies and related works done in this area. As the most challenging part of this thesis was the processing of the very large dataset, feature engineering was selected as the main focus of the thesis.

2.1. Data Mining

Data mining [2] is the process of extracting embedded data patterns by applying reasonable means of methods and techniques that form various data resources. The data sources may include data warehouses, databases, web, and data reservoirs. Mining is the term that refers to a digging process to collect a precious piece of nugget from a vast amount of ingredients. Data mining is similar to those processes, and it involves discovering those sets of data that can prove impactful from data resources. It has two significant approaches. First, it offers an overview of the collection of data to define and evaluate the main features. Secondly, the pattern detection focused on assessing the unusual patterns of behavior.

Data mining techniques are used in the financial, medical [3], scientific, research, and engineering field. In the medical field, data mining helps to explore the changes in folks DNA sequence influence the risk of creating diseases like tumors and cancer. In bio informatics, data mining is used in disease treatment, data cleansing, and gene finding. Micro-array technologies are used to promote patient outcomes. Data mining is used to tackle the business issues in the finance and banking sector by determining the casualties, correlations, and patterns in the market prices and business information. In the health care industry, data mining is used [4] to perform data analysis to identify primitive medical information. It is a tool that allows the decision-making process over banking, and its approaches are used for obtaining new customers and evaluating the customer's behavior.

In education [5], data mining is used for designing the analysis of data from the educational setting to understand students reactions. It helps to predict the future learning behavior of students. This approach is used by the institution to forecast the results of students and make an effective decision as well. Data mining enables us to analyze interconnected databases and find out queries with effective data mining techniques.Based on the vulnerability, the data mining technique allows identifying the segment of customers and improving satisfaction. Data mining operations use database analysis techniques to achieve high accuracy and efficiency when dealing with a vast amount of data sets. Database [6] systems evaluate multidimensional data mining by forming partially fragmented data cubes. Data mining has seen huge potential in many data-driven applications.

2.2. Machine Learning

Machine learning [7] is the recent technology used in various applications of day to day life. Significant factors like computational processing increased volumes, and a variety of data became capable and affordable data storage in this machine learning. It can automate the models which can be used to evaluate the complex data and generate accurate results and deliver faster. Organizations have a significance of determining the profitable opportunities and eliminating the unknown risks. Nowadays, companies are using machine learning to enhance productivity, business decisions, diagnose diseases and forecast weather, and so on. Due to the growth of this technology, we need to prepare ourselves for the data we have and need to understand the tools effectively.

Generally, machine learning has three primary types. In supervised learning, we contain both the input vector (training set) and output vectors (testing set). In unsupervised learning, we don't have any target vector. It is the more required form of learning due to no need for target values. The third and last type of machine learning is reinforcement learning, which refers to cases where we don't have any transparent input vector and output vector, and it usually learns from experience.

At present, machine learning applications could be found at every aspect of the modern world of information and technology from software, spam detection, medicine and healthcare, advertising, and robotics; all are heavily dependent on machine learning approaches. The objective of machine learning is used to evaluate the structure and patterns hidden in the data. These theoretical distributions are given to datasets to acquire a better understanding. It is a part of an emerging trend in companies that are using machine learning to solve real-world problems. Machine learning has proven its efficiency in tracking monetary fraud online.

Since the limitations of traditional marketing campaigns have been identified, there has been a need for such campaigns that are backed by strong concepts of machine learning so that the adverse effects caused by the limitations of the former, could be minimized. In addition to our work, there are a lot of studies that have been conducted to showcase the ability of smart marketing campaigns. In one such study, the synthetic data of customers of Starbucks [8], a famous coffee brand was used in modeling. The entire modeling procedure was split into two halves, modeling the consumer graphics and predicting the overall spending of the consumer. Later during

the execution process, the former was done using k-means clustering and the later, using XGBoost regression analysis.

PayPal [9] has employed machine learning for protection against money laundering. However, the company has set up a tool that helps them to map millions of transactions and differentiate the legitimate and illegitimate transactions which take place between sellers and buyers. An intelligent engine is a smart system that prefers equipment like RFID, sensors, Wi-Fi, and cellular communications links to receive data and interpret it to make decisions.

In another study, a prediction model has been developed, which can predict and then evaluate the number of ad impressions received by a marketing campaign [8]. The study aimed to evaluate the upcoming marketing campaigns by determining the best parameters, for instance, budget and timeline. Data from several hundred marketing campaigns was collected, and it was made sure that it had very few outliers. Secondly, the abundance of data implies more preciseness in the training of the model.

Four different supervised learning techniques were used in the mentioned study, namely, Decision Trees, Linear Regression, Random Forest, and Support Vector Regression. The training accuracy for all the mentioned techniques was recorded to be 0.67, 0.34, 0.75, and 0.63, respectively. The test accuracy for the techniques was recorded to be 0.64, 0.32, 0.66, and 0.61. The results of the mentioned study show clearly that the Random Forest classifier is the one that is giving the highest accuracy with both the testing and training data. The study also establishes the fact that with machine learning and data sciences, the traditional marketing paradigm can be revamped to a comparatively more effective and impact generating model. This study and the necessary implementation has been done to take the current digital marketing paradigm towards further effectiveness.

2.2.1. Random Forests

Predictive modeling structure consists of the following categories such as analyzed datasets of previous variables, features of ranking aspects of candidates, distinct classification of the datasets, generation of the feature for the candidate levels, evaluation and the models of selection techniques, achievement of approximately 25% reduction of the error values. Effectiveness of the future perspectives of engineering and machine learning categories is possible due to the improvement of the system architecture of random forest trees [10].

The well-developed architecture system of explore-Kit [11] is characterized as the improvement of the model of random trees and analyzing procedures. Various categories of algorithms are being used for proper description and analysis perspectives for a given problem statement. Meta-data techniques can also be used for better understanding. Meta-data consists of the following attributes, providing general information levels, evaluation at the stage of initial aspects, measures of entropy-based activities and analyzing procedures and performance of the methods of statistical tests and interpretation of the results. The diversity of future-oriented perspectives and techniques is possible by data collection through quantitative sources [12]. These data sources are the essential methods for accurate analysis of the random trees experiment procedures and techniques.

Random forest [13] is a meta estimator that matches the number of trees on different sub-samples of a dataset and enhances the accuracy in predictive nature. it is elementary and flexible to use that generate great result in most of the time without hyper-parameter tuning. It is suitable to be used for both classification and regression tasks, while it is the most common and used system concerning the features of its simplicity and diversity. In the machine learning process, the random forest algorithm is also identified as the random forest classifier. In general manner forest consists of trees, here tree mentioned decision trees. The random forest algorithms define the comparison of a random collection of a forest tree that identified as a decision tree algorithm. The main focus of the random tree is decision tree classifier, which correctly identified through entirely understand the nature of the random forest algorithm [14].

The random forest classifier is a system of decision tree forecasts and fixes the incidence to respond through voting. In the classification system, the collective voting of Decision Tree is the ultimate response, and in regression, that is the final response to assess all reactions. The assumption about the random forest [15] is an excellent process to train early in the model improvement method, to perceive how it implements. If anyone would like to develop a model very quickly, then it is more suitable for that. random forest is a generally fast, easy, and flexible tool, but has some limitations; regarding all of the random forests is the best and common models decision tree and forest classifier.

The generations of the future-oriented elements are essential for the characterization of MI (Machine Learning) processes. In this technique, proper analysis and collection of data sources are characterized by a random selection of the range of available data. It is impossible for the appropriate explanation of an explanatory model. The random forest can also be treated as a black box. Many steps are included for transaction systems for random forests and proper techniques [16]. Hence, future prediction is possible. An appropriate and effective sequence of events are characterized as the hyper-graph of different category of respondents with the techniques of data mining procedures.

The functional importance of Artificial Intelligence (AI) techniques includes these steps, such as prediction of the events in the future-oriented aspects, active participation of the co-entity levels, evaluation of the system of ranking points and understanding the influence of organizational structures, aggregation of the associated activities for a single analysis of a given graph and functional and accurate procedures of analysis techniques at the level-best efforts. For the categories of neural networks, it is vital for the study of neutral aspects. Improvement of the techniques of machine learning aspects is made effective by the use of computer networking systems and procedures.

2.3. Feature Engineering

Feature selection [17] is the process of selecting features that are most relevant when working with the predictive model. The objective is to find those properties from data that can predict the purchase decision. In this section feature selection or variable pattern recognition is used to reduce the number of attributes in the data sets and to create a new combination of correlative attributes. The method can for

example automatically add or remove attributes that exist in datasets. Allocating and distributing data can be time-consuming and requires effective computational determination. As machine learning algorithms have become more advanced and complicated, data sizes have also grown more abundant than ever.

Feature generation is the process of generating a completely new feature from the existing features to perform complex analysis operations. Not all features are relevant to perform predictive tasks. Therefore, selecting decorative, efficient, and independent features are fundamental of the utmost importance for the effective algorithm. Feature selection is a significant approach to reduce irrelevant dynamic attributes, improving learning accuracy and ensures results liability. Many feature selection methods have developed for machine learning applications.

In this section the **random variables** will be described because these are essential for analysis purposes. The reason behind the selection of random variables is characterized by the assigning of proper numerical value and favorable outcomes of the casual experiment activities. The probability of tossing a coin describes the importance of prospective studies. The significance of the random variables is essential for the selection of a given outcome. Hence the variety of random variables is crucial for the techniques of machine learning procedures. It can also be thought of as a function of favorable results for an experiment. For example, tossing a coin two times can lead to HH, TT, TH, HT, TT [18]. Also, the functional importance of choosing the random variables are or the prevention of cyber attacks.

According to [19], it has been stated that random variable is characterized as the function of probability with distinct aspects and improvement of studies of the performed experiment. It is also classified as the outcome to the quantity levels for the aspects of random variables. The random variable which is not continuous, are classified as the computational techniques of integration in the branch of mathematics.

For the categories of linear and exponential functions, the value of transformation is necessary. A straight line represents the character of the graph. The transformation graph is characterized by the importance of analysis of the regression aspects and the available linear functions [20]. Transformation graphs are widely used in the problems of algebraic expressions in mathematics. It is usually calculated on the XY plane. It is generally represented as

$$y = f(x) \tag{1}$$

where x is the vertical function, and y is the horizontal function. This kind of graph is mainly used for the transformations, and the effective modifications of the algebraic problems and proper solutions are being derived. This graph is helpful for the determination of the characteristics of a given function. Some specific importance is provided below for a better understanding. If the constraint of positive value(k) is applied, then the function will be

$$f(x) + k \tag{2}$$

then this graph possess upward characteristics. The values of k must denote the direction of the transformation graph. If the value of k is positive, then the graph will move leftwards and vice versa. Hence, the role of the transformation graph is important.

The range of the random variable is identified by the steps given below.

1. Listing of the events of the given sample space.
2. Determination of the probability for a particular event of sample categories.
3. Listing the values for possible values of X and the proper identification of the values.
4. Finding the events of simple aspects for which
$$X = k \tag{3}$$
where X is the pdf function and k is a constant.
5. Determination of the level of probability of equation 3 for the given size of simple event categories.

Also, the types of selection are emphasized by the critical aspects of continuous random variables and discrete random variables. Improvement of the process of selection is being made by the existing methods of re sampling methods, sampling by rejections, and the methods of transformation techniques [21]. It can be categorized by the proper techniques of machine learning processes and possible outcomes of a given experiment of the random aspects. Improvement of PMF (probability mass function) can also be used for the proper identification of random variables. The development of the gaussian theory is a useful approach for the identification of the random variable selections. The tossing of an unbiased coin is a good example.

Improvement of the decision tree is an essential factor for analysis purposes [22]. The efficacy of the random variables is characterized by the experiment of rolling a dice and tossing an unbiased coin. These two experiments have an expected outcome with different characteristics. The functional process of CFS techniques (Correlation Feature Selection) evaluates the magnitude of the subset values and the ability of the predictive aspects. Here, the process of genetic algorithms can also be used for better understanding using the step by step process for the solution of a given problem.

Data model is also used for the techniques of medical aspects. It is being used for the proper techniques of selection step for the cause and interpretation of valuable results. Importance's of the statistical aspects are characterized as the probabilistic categories of data collection tools and prospective techniques. The random variable of discrete values and continuous values are important for analyzing techniques, which can identify data sources for the techniques of analyzing purposes. Also, for the theory of probability it is important for the description of a specified event.

The random regression models are essential for the careful analysis of the sources of longitudinal data structures and the records which are repeated with aspects of time-oriented techniques. The functional importance of this model is for the study of interactions of the environment and the variability of the genetic elements. A normal distribution characterizes the level of residual error, and having a value of expectation is 0, with variance level is sigma e square. For the proper techniques of sampling methods, it is denoted as Var (X). It is widely used in the statistics of the descriptive aspects [23].

The linear regressions [24] are the characteristics of a description of the relationship between dependent and independent variables. The functional importance of linear regression is used to determine the level of predictions or information forecasting techniques. By this process, the reduction of errors is also possible. For the critical aspects of machine learning processes, understanding of the functional algorithm is also essential. Thus, accurate predictions are likely due to proper procedures of analysis and testing methods. Furthermore, linear regression is used to compare the different characteristics such as height and weight, and so on. Improvement of data collection is possible due to assumption techniques and gaussian distributions, and so on.

Data analytic bears prime importance for many industries due to the availability of massive databases, open-source machine learning tools, and cheap computational resources. Feature engineering involves understanding of domain knowledge and data exploration to discover relevant features from the raw database, which would subsequently be used by machine learning tools for analysis [25]. Automation of the feature engineering is intricate yet crucial because it would speed up and enhance the effective use of machine learning models. Significant challenges encountered in developing a fully automated feature engineering systems involves diverse basic data types, their complexities, temporal variations, complex relational graphs, and large transformation search space requirements.

In this section **Deep feature synthesis** (DFS) is described because it is a process of automated feature generation. It is obtained from the relationships between data points within a dataset such that it performs feature engineering for transactional datasets that are in most cases acquired from log files and databases. This type of data is the key focus as it is the most prevalent enterprise data used in modern times. According to [26], a research on 16000 data scientists realized that over 65% of their time was spent handling relational datasets.

It is also notable that any new features within the deep feature synthesis is acquired from previous features. Consequently, the DFS is made up of primitives, which define both input and output types. They can, therefore, be stacked on each other to create complex features that ape those established by humans today. DFS is applicable to relationships between entities to establish new features from existing datasets while controlling their complexity via establishing a maximum depth in the optimized search.

There are several advantages associated with the DFS process [27]. These advantages are functional relationships between the data entities, and a single set of data structured are developed—proper synthesizing of the required sources of data within the different subsets of available data sources. Adequate and valid identification of relationships of the various entities help the derivation of new features and analyzing purposes. Improvement of the process of machine learning aspects is developed effectively. The technique can also improve overall developments of digitization aspects that are enhanced with the scaling of the infrastructure of data and perspective theories. The DFS process is especially handy in the analysis and interpretation of results.

In this section **random forest** is used for feature induction [28]. The research suggests a simple yet effective approach to induce a task-dependent feature demonstration using ensembles of random decision trees. They find the benefit of that learning directly a task-dependent feature representation rather than learning a kernel

function and review the kernel learning methods. Here are methods that attempt to learn a linear transformation of the original kernel, tackling the so-called mahalanobis learning task or that learns a linear combination of kernels belonging to a specific family of parameterized functions or multiple kernels approaches. The method is inductive, fully non-parametric, and can be informed by a wide range of supervised circumstances. They accumulate all the features of every decision tree in a random forest through a hashing system to get the final encoding. He uses the random forest version, as suggested by Breiman [29], that combines bagging and random feature selection and apply the bootstrap procedure to create multiple randomized training sets.

He introduced a simple non-parametric method that uses the structure of an ensemble of random decision trees to derive a task-dependent feature encoding from investigating the characteristics in the experiment. He target alignment and classification performance of kNN and SVM classifiers built on top of such illustrations recommend that the proposed procedure is effective in integrating information from a given specific task. They design induced features to get the better predictive performance and the data using relational random forest learners to feature the construction process.

In this section the support vector machine has used in email spam classification because of a higher number of email datasets and features [30]. To enhance the support vector machine, improve the classification accuracy, and minimize computational complexity. Feature selection is highly dependent upon the ANOVA f-test statistics scheme was used to evaluate the significant features for email spam classification. The feature selection used to minimize the high data dimensionality of feature space based on one way ANOVA F-test before the classification process. The proposed scheme was determined by using a spam base benchmark dataset to assess the feasibility of the proposed method. This comparison has been used for categorization algorithms, success measures, and different datasets. The experimental results on spam based indicated that improved SVM manages SVM and many spam classification methods for the english dataset concerning dimension reduction and computational complexity. The feature selection method based on ANOVA F-test is used to escalate the insignificant attributes from the dataset. The proposed scheme of study reduced 57 features of the spam dataset to 52 to avoid the unrelated feature to eliminate the low classification accuracy and high data dimensionality using ANOVA.

The feature selection focus on identifying the best subset comprises m features extract from n features. The classifier system is divided into two parts, namely testing and training. In this study, SVM used as classifiers and chose key data points as it supports vectors for prediction. The scheme has advantages such as minimize the false positive rate and computational time and enhanced classification accuracy.The time cost is 63.09 sec, and the false-positive is 0.04, and the classification accuracy is 93.5%. The result of the new scheme was compared to spam detection using SVM and offers better computational time, classification accuracy, and false-positive rate.

There are some adopted approaches to evaluate the spam predictors like t statistic measure of significance between two means of spam and non-spam subsets of data. The performance was compared to some other classifier algorithms using an email dataset. This study deploys the feature selection with an evolutionary algorithm to offer useful and accurate results. It is used to evaluate the optimal features. The new spam

detection is depending upon the hybrid between ANOVA as a feature selector and SVM based on the poly kernel as a classifier. The integrated method could attain effective results for computational time and classification accuracy on the spam base dataset.The proposed method acquired good results by enhancing efficiency and minimize the time cost compared to spam detection. The accuracy score has ANOVA F-test as a feature selector to SVM.

The researcher addressed that high dimensionality of thousands of features and micro array [31] data in some samples are considered to be a challenge that affects the analytical results. The support vector machine is used for the classification of micro array datasets, but the high dimensionality of feature space issues still exists. In this study, the authors highlight the minimization of gene expression data into a minimal subset of genes with the help of feature selection. In order to minimize the computational burden and noise emerging from irrelevant genes, machine learning is used in the classification of cancer from micro array data. Many machine learning algorithms and statistical theory choose essential features and eliminate irrelevant and redundant features. However, it is not clear how these methods and algorithms react to small sample sizes. It offers the combination of ANOVA for feature selection and minimizes high data dimensionality of feature space and reduces computational effectiveness and complexity using support vector machine algorithms. In addition to this, the algorithm eliminates the computational noise and burden emerging from irrelevant and redundant features. It also minimizes the gene expression data to a low number than thousands of genes, which can eliminate cancer testing costs. It uses a colon cancer dataset that comprises DNA micro array gene expression data with 62 samples and 2000 features.

In this section analysis of variance (ANOVA) is used for feature selection. The proposed approach chooses the informative subset of features for classification to acquire high sensitivity, accuracy, precision, and specificity. The feature selection approach in micro array technology serves a significant role in forecasting and treating diseases in medical research. The feature selection method is used to identify notable features. By using one way ANOVA method, the classification accuracy rate has achieved, and the SVM classifier is 86.67%. The features are minimized from 2000 to 416 features. The experimental results implicate that the proposed method has effective performance compared to SVM methods as we know ANOVA is the best ranking method for identifying the smallest gene subsets to attain the perfect cancer classification. SVM can enhance the performance of generalization by accurate selection of kernel to a particular application. The results evaluated from the colon dataset for gene combination offers classification for uniqueness. It is used for the feature selection method to manage the treatment effect and choose essential genes that enhance the performance of classification.

It is a tree regularization [32] framework to accelerates several tree models for feature selection efficiently. The focus is to penalize selecting a new feature for splitting and avoiding the earlier structure. The experiment represents hypothetical assumption and defines variables how it applied and multi-dimensional outcomes of the framework for finding a solution of feature selection. The researcher leads its study by defining some examples to select target variables like X1, X2 to set depth indicators for outcomes and include a sample of the study. The tree model deals with categorical

and mathematical variables, omitted values, different scales between variables, and interactions to the regularization framework.

It always highlighted different models of feature selection for picking the best subset to bring about the learner. It defines a famous example that is the SVM-RFE model. To the proper analyzing solution, take several models to interact variables and optimizing a decision tree, which avoids feature redundancy. They also tried to figure out a relation between the decision tree and the max-dependency scheme, also represent the regression of a different angle to define theoretical and mathematical calculations over the study.

It theoretically measures the evaluation and performance of the feature selection method. To maintain minimal feature set without changing predictive information, he mentions about markov blanket of Y. He also noted empirical measure to set of training instances of instantiating of feature set X drawn from distribution D for proper understanding. The researcher defines a statistical experiment to combine the average accuracy of different algorithms and using the feature subsets. He focuses on engendering high-quality feature subsets for both weak and robust category. It also tries to find out a mathematical solution to interact with several variables for effective and efficient feature selection.

In this study, we are describing different feature selection methods for the Optimal design [33] of studies for developmental inquiry. The researcher suggests the use of feature selection, a data-driven machine learning algorithm in the study proposal, and a variety of measures that demonstration the most predictive power in pilot data. They illustrate how data-mining systems, precisely feature selection methods, might be used to optimize item selection to maximize prediction with a summarized assessment protocol. Here proposed a method of feature selection that categorized in two analytical tools: feature importance and recursive feature elimination. Here include an example that feature importance in a regression model might be calculated in terms of the t value used to subtract the predictor's level of numerical significance by Bursac, Gauss, Williams, & Hosmer [34]. They use the method to plan a monthly study of self-reported health and life satisfaction by choosing a set of measures from the existing SOEP data.

He uses an automatic model selection process to reduce the number of variables before calculating feature significance. They also suggest RFE selects a model for each outcome with a number of measures and more consistent results with DTEs or support vector machines. And the most important and decision making step is to select features. This is evaluation part of using fitness measure through RFE to ensure that the final subset maintains reasonable predictive accuracy after manual selection.in the stage of result they categorized a subgroup of 20 measures from the SOEP data set that continue much of the ability of the original data set to calculate life satisfaction and health across the younger, middle, and older age groups through using the methods. The study focuses on feature selection approaches that make it the potential to combine data-driven insights with functional and practical expertise to select an optimal set of predictors for a study or analysis.

This research represented a unique numerical technique through permutation tests for tree-based [35] supervised learning methods regarding variables set. It examines both traditional and new ways to calculate false discovery rates of the uni variate variable for selecting multivariate tree-based importance variables. In the context of

the study, researchers suggested two balancing sub-problem for the feature section. The first sub-problem indicates relevant variables to customize the target output of the research and the second sub-problem classifies about proposed set of variables that do not define any target variables. They consider problem selecting relevant features through probable selectivity and compromise subset of relevant variables sustaining the rate of false positive. They partially used statistical analysis for some specified questions to seek target output. First of all, they asses a false directory rate (FDR) to measure tree-based multivariate importance. They mentioned a method of real FDR how to lead random selections of relevant subsets. In this research, they have taken a study sample of N input-output sets from some undefined perspective. The m input variables symbolized.

$$f_i, (i = 1, \ldots, m) \tag{4}$$

The experiment focused on tree-based methods, the idea of classification trees comes from [36]. Here suggested two tree-based ensemble methods that are based on randomization, namely random forests and extra-trees. There are some specified variable importance measures suggested for tree-based methods. They mentioned, information theory for examining node n total decrease through,

$$I(n) = \#S,H_C(S) - \#S_t,H_C(S_t) - \#S_f,H_C(S_f) \tag{5}$$

To authenticate the proposed methods, researchers create a pair of artificial problems so that they can check relevant variables correctly. They used the matlab code to generate a dataset for the NIPS 2003 feature selection challenge. First, they take a 3-20 dataset to collect 200 objects and 20 variables. Where the first three are relevant and the rest of them gaussian noise. And the second dataset takes 50-1000 samples for managing 2000 objects and 1000 variables. Writers take a learning sample to calculate tree-based method. To select a subset of variables, aggregate the value of the threshold, accepting false positive to choose the variables. researchers pretend a null hypothesis for executing the tree-based method, which used to generate importance on datasets. Writers also suggested an alternative measure to relate with significant threshold for defining variables that are ranked.

They proposed this method of random computing permutation so that the value of variables remains unchanged. They also experiment with simulated data through conditional error rate (CER) and FDR regression. Here focuses the measure on the real problem of testing biological datasets. They used a couple of numerical processes for multiple uni variate statistical tests that are relevant to variables. The whole study writers represent two statistical methods to extract a subset to relevant variables through tree-based supervised learning methods, which are FDR for uni variate statistics testing outline and CER as a new method for feature selection. They also used machine learning algorithms for multivariate importance measures regarding tree-based practices.

ExploreKit [11] is a framework that is used for generating and selecting features automatically. It is a repetitive process where each repetition consists of three steps. In the first step, it creates a broad set of candidate features from the given set of features by using some arithmetic operators like unary operators, binary operators, and higher-

order operators. In the next step, that is candidate features ranking in which he allocates a score to every candidate feature that totally depends on its predictable contribution. In the last step, he used a greedy search method to evaluate the ranked candidate features. The experiments are performed on 25 different supervised classification datasets. We used three classifiers for evaluation.

It is about the **probabilistic approach** [37] to feature selection (a filter solution). According to the researcher, feature selection is the process under which we find minimum M relevant attributes of a dataset while considering its original total number of characteristics. Feature selection is mainly categorized into two main categories of wrapper approach and filter approach. However, the filter approach can be further divided into two groups, known as heuristic search and exhaustive search. In the article, the researcher has discussed the key advantages of all of these approaches in real-life implementation.

According to the research findings, the critical advantage of feature selection is its ability to reduce dimensionality in the datasets. Moreover, LVF is the most effective approach to conduct a test for the real massive datasets. Although, it is beneficial for the confidential datasets to the implemented machine learning algorithm. Additionally, feature selection by the use of the filter approach provides a straightforward method to replace inconsistency and accuracy issues regarding learning algorithms in various datasets. It can also eliminate the chances of doubt in the running time of LVF.

Based on the research analysis, LVF is the best choice that should be given preferences in feature selection and dataset analysis (with massive datasets). Summarizing the whole discussion about feature selection, researchers concluded in the end that the probabilistic approach to feature selection offers two quite different methods for contracting a conventional analysis of datasets. The theoretical analysis represents that a filter solution can be made possible in a dataset when all optimal resources are permitted to use, and selector requires inconsistency criterion. Besides, the probabilistic approach is identified as a fast and guaranteed approach.

We can select new variables by using a **random forest classifier** [38]. The research vastly describes Random Forests, Variable selection, Permutation test, and multiple testing using several techniques. They define Random forest as a frequently applied algorithm for classification and regression problems and get a higher accuracy rate of prediction. The CART algorithm explains the random forest of recursive partitioning. In the process of variables selection, they mentioned different research fields the application of cross-validation and summarized the discussion of selection with random forests and also explain performance-based approaches and test-based approaches. The proposed new method of variable selection for two reasons to develop the accuracy of a prediction model and identify the most relevant and informative variables from a set of candidates. This study enlarged the selection idea and initiated new approaches of variables selection using random reforest and the permutation test framework to run the new system, firstly makes it possible to control the TWER and FWER, Secondly, a higher power to distinguish relevant from irrelevant variables compared to conventional methods and lastly to achieved the highest ratio of related to selected variables. They believe that new suggested approaches improved the accuracy rate of selection in the random forest.

Feature selection [39] (FS) is a popular topic owing to its applications in several data sciences related studies. The approaches of feature selection including,

forward selection Heuristics and backward selection heuristics, meta-heuristics, and evolutionary algorithms, are being employed in lots of research areas. In a study whose primary focus was to search the space of alternative feature subsets efficiently, and to provide the ranking of specific attributes in accordance to their relevance for the prediction task, feature selection played a pivotal role by successfully predicting those such features using which; a model was capable of identifying top 20% of the attributes which could yield 85% of the accuracy.

Recently, data mining and machine learning-based approaches have been widely applied by several contemporary enterprises solely to make an effort to improve the retention rate of the customers. In one such study, a novel approach has been introduced, which can predict the repurchasing demand of a group of small & medium enterprises (SMEs) customers in a large-scale online recruitment firm [40]. Two feature selection techniques have been employed in the respective study; namely, correlation based feature selection (CFS) and subset consistency feature selection (SCFC). These two approaches were used to develop a predictive model, and the results were compared with those models in which feature selection was not used. The results that were produced were analyzed comparatively, and it was established that the model with feature selection techniques performed remarkably than their non-feature selection based counterparts. It further gave rise to the theory that a predictive model based on feature selection can perform better.

3. SYSTEM DESIGN AND IMPLEMENTATION

3.1. Data Description

In this thesis, the data collection for each user in the ambiance of a video racing game, which was launched by a well recognized company. It is a modest boundless car driving game that requires the user to drive a car on a hilly territory avoids collapsing or breaking the neck. The new game is not just visually comfortable but also marks a shift from the endless driving category to the racing genre. It has interactive game-play and design with eye-catching graphics in average package size. It is all about a 2D perspective with a unique approach that you can easily filter the mode on your smartphone. The version is quite similar to the first edition, but it has included several new features and options. It added customize gaming mode and changes the character as you like. You can quickly get a high score while driving the car and add a library of new vehicles with some stages that you have to upgrade for a better experience. There is a tutorial and user-friendly interface for the new player if you face any difficulty then check this option. There is also a purchase option for the stage that helps to unlock new stages and options.

The most vital motions are to play the game and finish races where players would like to pick a car from multiple options. Completing a race, you can get coins and rewards, then the levels of movement of the game are shown as different racing stages that hold coins to collect. If you would like to go to a certain level of the game, you have to title the car that is a mix of forwards and backward directions to move your vehicle much more unpredictable and make a preference to control the unit. This game is more sequential to other video games on the ground of control. If you would like to achieve an extreme range of nods like a back-flip, front flip, or even multiple aerobatic motions, you need to keep the accelerator pedal pressed for reaching a front flip. There is some approach to this game, the two main modes of them, such as endurance type, known as adventure, and defeating AI drivers. In this game, each stage has its difficulty curve that upturns as you unlock and purchase the highest stages and also include the practice section to overcome the difficulty. There are also several features such as weekly events open, tune, team up, explore, social media connection, customize menu, and compete to climb the leader-board.

It is all about the presence top on the leader-board that can ensue when you have an acute sense of the things needed to get the highest score. To get the highest score and reach a certain level of the game, you have to be more patient and practical; categories are dividing the algorithms to boost your score and overcome top-level, such aerial aerobatics, repeating levels and collecting the big coins. If you would like to purchase in a single stage to keep progress, the machine will recommend you upgrade the best time of purchasing while you find the difficulty. Where you feel trouble climbing over the steep hill, which means you need more powerful engines; here, you can use coins to upgrade engine. In the same way, if you face difficulty in landing after big air time, your car has to better suspension and grip; upgrading the grip improves high flying off-speed and covers more distance in a shorter time. Customization and upgrade parts can be purchased as coins or gems with your own currency as you want.

According to the Firebase console, data is from the last month, the user demographics are varied. The figure shows that 72.5% of users are male, and 27.5%

of users are female. In the figure, the age period between 18 to 24 has 24.95% of the male of the total users. The age period between 25 to 34 has 9.94% of females of the total users. There are 15.8% of users from India, 13.6% of users are from the United States, and 8.4% of users from Russia. These three countries show the most significant amount that is 37.8% of total users.

The experiments we did in this study were performed with the game repository. We have chosen the six days of data from the daily_user_state dataset and daily_sessions dataset from the repository. The daily_user_state dataset consists of 35 attributes and 2 million instances. The daily_sessions dataset has 11 attributes and 2 million instances. Both datasets have some nominal attributes and some of them are numeric.

3.2. Design

To evaluate the data, we need to design the software system. In this thesis, machine learning techniques were used to classify the data specifically we have used random forest classifier [29] because it has a high accuracy rate, considering that it uses four to twelve hundred decision trees in the feature generation. Each of the trees is generated over a casual observation abstraction from the dataset. Features are extracted randomly, which makes the random forest have no over-fitting. Even though not every tree can see all the features or observations. Each tree also combines yes or no queries derived

Features name	Type	Description
rew_ad_revenue	FLOAT	revenue that generates by watching ads
rew_ads_watched	INTEGER	Number of ads that tells about in-game rewards watched in the current day
records	INTEGER	Number of records in the game in the current day
daily_gem_state	INTEGER	Amount of gems that user has in the current day
sess_length_seconds	FLOAT	Total sum length in seconds of the sessions during the current day
daily_max_rank	INTEGER	Maximum game rank in the current day
ads_watched	INTEGER	Number of ads that user watched
total_sessions	INTEGER	Number of sessions that user used in the current day
daily_min_rank	INTEGER	Minimum game rank in the current day
daily_coin_state	INTEGER	Amount of coins that user has in the current day

Table 1. Feature Set with Description

from a single or combined set of features. It eliminates the over-fitting that is a useful element in feature generation and makes the approach even more useful. This process of applying machine learning concepts is to transform the information into knowledge. It helps to determine the structure in a large amount of data. It will contribute to decision patterns in the given data. In addition to this, I used the data preprocessing approach for transforming the raw data into a compelling format. Figure 2 shows the steps that we followed for selecting the relevant features.

3.3. Implementation

This part will describe the implementation of my work. I used machine learning because it is a system of pattern recognition to generate reliable and informed results from the data, and it is a core sub-area of artificial intelligence. Machine learning can quickly drive difficult tasks and get better output to analyze vast datasets such as web search results, web pages, mobile devices, email spam filtering, network interference detection, and pattern recognition.

3.3.1. Data Collection

At the beginning of this task, I would like to explain existing data on the recognized IT company named Fingersoft that has a long history of this field and maintain a large amount of data. The organization prefers to use several data processing tools concerning the present situation and available data. The primary source of data was the google big query, which preserving through the firm. The process of data interpretation goes the part of data science that they store vast data in a structured format. In this step, the main task was to establish a connection of google big query with python to use the external data sources. The data that I used in experiments was the earlier days data of game. I used two entities named daily_user_state and daily_sessions as shown in Figure 3 and Figure 4

3.3.2. Preprocessing

In this section, I will discuss the preprocessing of the data and explains the different techniques that I use in this work—preprocessing mentions to the transformations useful to our data before nourishing it to the algorithm [41]. It is a method that used to transform the raw data into a clean data set and indifferent term; when the data is grouped from diverse sources, it is collected in raw format, which is not feasible for

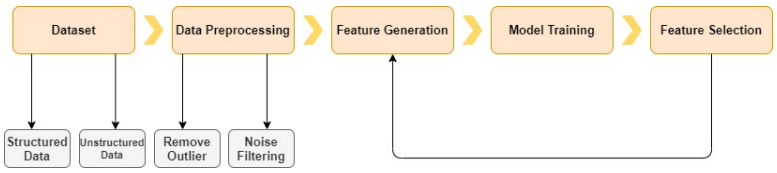

Figure 2. Design Overview

the analysis. The data that I am using has different issues like it was incomplete NaN values, and there is also an issue of class balancing. So I use the data preprocessing techniques of resolving such matters. The stage of data preprocessing was very critical, and there are several steps of it like import the libraries, import the dataset, check out the missing values, perceive the categorical values, splitting the dataset into training and test Set. I

These above steps of data preprocessing mainly polish data differently and make data use-able for the classifier. There was another issue of class balancing in the data because I have two label classes in the data that were yes and no. It was a significant challenge for classification models is an imbalance of classes in the training data. I used earlier days data of the racing game, so I have like five million false labels and twenty thousand true labels where a customer is purchasing. So, I have only 0.4 percent values of true label and 99.6 percent values of the false label. There was a significant class imbalance issue, so I proposed a new design to resolve this problem of class imbalance. I took 20000 yes label values and randomly chose 20000 no label values, and in this way, I solved the issue of class balancing.

creation_date	activity_date	platform	app_version	country	country_code	retained_days	purchases	purchases_valid	revenue	net_revenue	int_ads_watched
2020-04-25	2020-06-08	android	1.36.6	Guatemala	GT	44	0	0	0.0	0.0	11
2020-05-31	2020-06-08	android	1.36.7	Chile	CL	8	0	0	0.0	0.0	10
2018-11-28	2020-06-08	android	1.36.7	Belgium	BE	558	0	0	0.0	0.0	2
2020-02-29	2020-06-08	android	1.36.6	Taiwan	TW	100	0	0	0.0	0.0	10
2020-05-30	2020-06-07	android	1.36.7	Venezuela	VE	8	0	0	0.0	0.0	16
2020-05-23	2020-06-08	android	1.36.7	Australia	AU	16	0	0	0.0	0.0	0
2020-05-10	2020-06-08	android	1.36.7	North Macedonia	MK	29	0	0	0.0	0.0	7
2017-08-02	2020-06-08	ios	1.36.5	Brazil	BR	1041	0	0	0.0	0.0	1
2020-06-06	2020-06-08	android	1.36.7	Vietnam	VN	2	0	0	0.0	0.0	11
2020-06-07	2020-06-08	ios	1.36.5	Australia	AU	1	0	0	0.0	0.0	24

int_ad_cost	rew_ads_watched	rew_ad_revenue	rew_ad_cost	ads_watched	ad_revenue	daily_min_rank	daily_max_rank
0.00032329126992304623	9	0.0068304090112683343	0.00075893433458537043	20	0.010386612980421843	34033	37273
0.00051714726792162014	15	0.022788238457456342	0.0015192158971637562	25	0.027959711136672543	23742	25185
0.0012355661409561	21	0.061228567085154277	0.0029156460516740132	23	0.0636996803133455	914079	916539
0.0011402063659786742	32	0.24349000325407483	0.0076090626016898383	42	0.25489206691386157	810000	813540
0.00011199385306463065	7	0.0026222347868123121	0.00037460496954461604	23	0.0044141364358464023	6589	14368
0.0046800553230577366	33	0.31694490576866369	0.0096043910838989	33	0.31694490576866369	33550	36110
0.00017499423213455511	15	0.005608664020141675	0.00037391093467611166	22	0.0068336236450835608	56993	57958
0.0015557900740756445	22	0.086026953020030286	0.003910316046365013	23	0.087582743094105936	1731280	1733621
0.00048527951125207385	20	0.027419337491203618	0.0013709668745601809	31	0.032757412114976431	661739	665279
0.0054612509444703578	10	0.11657454652217859	0.011657454652217859	34	0.2476445691894672	0	14444

daily_coin_state	daily_gem_state	garage_power	mobile_brand_name	mobile_model_name	mobile_marketing_name	daily_min_season_rank	daily_max_season_rank
13537	3	751	*null*	*null*	*null*	0	0
11849	11	201	Samsung	SM-J320M	Galaxy J3 Duos	0	0
43033	3	2024	Samsung	SM-A600FN	Galaxy A6	31	33
1244161	428	3672	OPPO	CPH1719	R11s Dual Sim	26	30
4634	3	212	Samsung	GT-I9190	Galaxy S4 Mini	0	0
49864	348	827	Samsung	SM-G960F	Galaxy S9	0	0
43396	57	754	Samsung	SM-A307FN	Galaxy A30s	5	6
1803950	15138	4980	Apple	iPad Air 2	iPad Air 2	12	14
51530	738	5701	Samsung	SM-G610F	Galaxy J7 Prime	2	6
8423	45	154	Apple	iPhone X	iPhone X	0	0

Figure 3. Some samples of table daily_user_state from the game database. All the three blocks represented by the square brackets show different features for the same samples.

sess_id	country	platform	version	activity_date	sess_start	sess_end	records	total_sessions	sess_length_seconds
1	India	android	1.36.7	2020-06-08 01:46:41.814 UTC	1591580801814000	1591582495234001	219	4	1693.4
3	Romania	android	1.36.7	2020-06-08 07:25:52.146006 UTC	1591601152146006	1591601165087002	7	8	12.9
1	Kazakhstan	android	1.35.0	2020-06-08 16:15:05.797 UTC	1591632905797000	1591632934659002	3	1	28.9
1	India	android	1.36.7	2020-06-08 07:24:09.653 UTC	1591601049653000	1591601551689010	98	2	502.0
5	United States	ios	1.36.5	2020-06-08 18:43:00.703 UTC	1591641780703000	1591642373303040	117	9	592.6
2	India	android	1.36.7	2020-06-08 10:07:08.372 UTC	1591610828372000	1591610831696005	9	2	3.3
1	Germany	ios	1.36.5	2020-06-08 04:11:46.917 UTC	1591589506917000	1591590393353001	291	9	886.4
6	Spain	android	1.36.7	2020-06-08 08:28:49.255 UTC	1591604929255000	1591606640505001	157	25	1711.3
1	Czechia	android	1.36.7	2020-06-08 12:25:55.871 UTC	1591619155871000	1591619728306019	135	2	572.4
13	Canada	ios	1.36.5	2020-06-08 20:52:52.105 UTC	1591649572105000	1591649917126002	55	13	345.0
1	United States	ios	1.36.5	2020-06-07 21:00:01.626 UTC	1591563601626000	1591565208525000	229	7	1606.9
3	Finland	android	1.36.7	2020-06-08 14:45:51.391 UTC	1591627551391000	1591629639851010	308	5	2088.5
4	South Korea	ios	1.36.5	2020-06-08 08:08:18.988 UTC	1591603698988000	1591603700506000	10	6	1.5
1	Uruguay	android	1.36.7	2020-06-07 21:24:59.597 UTC	1591565099597000	1591565574073001	71	14	474.5
13	India	android	1.36.7	2020-06-08 06:02:51.517 UTC	1591596171517000	1591596325222001	22	39	153.7
9	Taiwan	ios	1.36.5	2020-06-08 11:45:45.526 UTC	1591616745526000	1591616813302001	23	10	67.8
1	India	android	1.36.7	2020-06-08 07:13:19.739 UTC	1591600399739000	1591601467677007	213	4	1067.9
1	Poland	android	1.36.6	2020-06-08 04:11:41.777 UTC	1591589501777000	1591590937822014	225	7	1436.0
2	Poland	android	1.36.6	2020-06-08 04:41:31.591011 UTC	1591591291591011	1591594984430011	476	7	3692.8
1	Ukraine	android	1.36.7	2020-06-08 17:30:40.161 UTC	1591637440161000	1591637709916006	73	1	269.8
2	United Kingdom	android	1.36.8	2020-06-07 22:03:43.087 UTC	1591567423087000	1591567944700011	66	14	521.6

Figure 4. Some samples of table daily_sessions from the game database.

3.3.3. Feature Generation

In this section, we will generate the features from the feature set. In feature generation, we produce different features from one or multiple existing features, theoretically for use in numerical analysis. The system increases new information to be manageable through the model construction and thus hopefully results in more accurate and useable. We generate some features by analyzing the feature set in figure 5 and 6. We know that the whole machine learning processes use some input data to create outputs, and this input data encompass features, which are generally in the form of structured columns. Features generation is preparing the accurate input dataset that is compatible with machine learning algorithm requirements and improving the performance of machine learning models.

3.3.4. Model Training

Model selection is the phenomenon of selecting one from various models for predictive modeling [42]. Some concerns occurred while evaluating the model selection over model performance like maintainability, complexity, and availability. The model selection needs adequate data, which can be based on the problem complexity. The model selections are classified into test sets, validation, and training, which meets the candidate models on the training sets. It can determine the set and report the performance of the test set. The training set is used to meet the model; the test set is used for managing the generalization error of the selected model, and the validation set is used for evaluating the prediction error for model selection.

The machine learning model is defined as the model artifact, which can be developed by the training process. There will be some issues that emerge from predictive

creation_date	activity_date	platform	app_version	country	country_code	retained_days	purchases	purchases_valid	revenue	net_revenue	int_ads_watched
2020-04-25	2020-06-08	android	1.36.6	Guatemala	GT	44	0	0	0.0	0.0	11
2020-05-31	2020-06-08	android	1.36.7	Chile	CL	8	0	0	0.0	0.0	10
2018-11-28	2020-06-08	android	1.36.7	Belgium	BE	558	0	0	0.0	0.0	2
2020-02-29	2020-06-08	android	1.36.6	Taiwan	TW	100	0	0	0.0	0.0	10
2020-05-30	2020-06-07	android	1.36.7	Venezuela	VE	8	0	0	0.0	0.0	16
2020-05-23	2020-06-08	android	1.36.7	Australia	AU	16	0	0	0.0	0.0	0
2020-05-10	2020-06-08	android	1.36.7	North Macedonia	MK	29	0	0	0.0	0.0	7
2017-08-02	2020-06-08	ios	1.36.5	Brazil	BR	1041	0	0	0.0	0.0	1
2020-06-06	2020-06-08	android	1.36.7	Vietnam	VN	2	0	0	0.0	0.0	11
2020-06-07	2020-06-08	ios	1.36.5	Australia	AU	1	0	0	0.0	0.0	24

int_ad_cost	rew_ads_watched	rew_ad_revenue	rew_ad_cost	ads_watched	ad_revenue	daily_min_rank	daily_max_rank
0.00032329126992304623	9	0.0068304090112683343	0.00075893433458537043	20	0.010386612980421843	34033	37273
0.00051714726792162014	15	0.022788238457456342	0.0015192158971637562	25	0.027959711136672543	23742	25185
0.00123555661409561	21	0.061228567085154277	0.0029156460516740132	23	0.0636996803133455	914079	916539
0.0011402063659786742	32	0.2434900032540748 3	0.0076090626016898383	42	0.25489206691386157	810000	813540
0.00011199385306463065	7	0.0026222347868123121	0.00037460496954461604	23	0.0044141364358464023	6589	14368
0.0046800553230577366	33	0.31694490576866369	0.0096043910838989	33	0.31694490576866369	33550	36110
0.00017499423213455511	15	0.005608664020141675	0.00037391093467611166	22	0.0068336236450835608	56993	57958
0.0015557900740756445	22	0.086026953020030286	0.003910316046365013	23	0.08758274309410 5936	1731280	1733621
0.00048527951125207385	20	0.027419337491203618	0.0013709668745601809	31	0.032757412114976431	661739	665279
0.0054612509444703578	10	0.11657454652217859	0.011657454652217859	34	0.2476445691894672	0	14444

daily_coin_state	daily_gem_state	garage_power	mobile_brand_name	mobile_model_name	mobile_marketing_name	daily_min_season_rank	daily_max_season_rank
13537	3	751	*null*	*null*	*null*	0	0
11849	11	201	Samsung	SM-J320M	Galaxy J3 Duos	0	0
43033	3	2024	Samsung	SM-A600FN	Galaxy A6	31	33
1244161	428	3672	OPPO	CPH1719	R11s Dual Sim	26	30
4634	3	212	Samsung	GT-I9190	Galaxy S4 Mini	0	0
49864	348	827	Samsung	SM-G960F	Galaxy S9	0	0
43396	57	754	Samsung	SM-A307FN	Galaxy A30s	5	6
1803950	15138	4980	Apple	iPad Air 2	iPad Air 2	12	14
51530	738	5701	Samsung	SM-G610F	Galaxy J7 Prime	2	6
8423	45	154	Apple	iPhone X	iPhone X	0	0

Figure 5. Generated feature from daily_user_state table are in black rounded rectangle

sess_id	country	platform	version	activity_date	sess_start	sess_end	records	total_sessions	sess_length_seconds
1	India	android	1.36.7	2020-06-08 01:46:41.814 UTC	1591580801814000	1591582495234001	219	4	1693.4
3	Romania	android	1.36.7	2020-06-08 07:25:52.146006 UTC	1591601152146006	1591601165087002	7	8	12.9
1	Kazakhstan	android	1.35.0	2020-06-08 16:15:05.797 UTC	1591632905797000	1591632934659002	3	1	28.9
1	India	android	1.36.7	2020-06-08 07:24:09.653 UTC	1591601049653000	1591601551689010	98	2	502.0
5	United States	ios	1.36.5	2020-06-08 18:43:00.703 UTC	1591641780703000	1591642373303040	117	9	592.6
2	India	android	1.36.7	2020-06-08 10:07:08.372 UTC	1591610828372000	1591610831696005	9	2	3.3
1	Germany	ios	1.36.5	2020-06-08 04:11:46.917 UTC	1591589506917000	1591590393353001	291	9	886.4
6	Spain	android	1.36.7	2020-06-08 08:28:49.255 UTC	1591604929255000	1591606640505001	157	25	1711.3
1	Czechia	android	1.36.7	2020-06-08 12:25:55.871 UTC	1591619155871000	1591619728306019	135	2	572.4
13	Canada	ios	1.36.5	2020-06-08 20:52:52.105 UTC	1591649572105000	1591649917126002	55	13	345.0
1	United States	ios	1.36.5	2020-06-07 21:00:01.626 UTC	1591563601626000	1591565208525000	229	7	1606.9
3	Finland	android	1.36.7	2020-06-08 14:45:51.391 UTC	1591627551391000	1591629639851010	308	5	2088.5
4	South Korea	ios	1.36.5	2020-06-08 08:08:18.988 UTC	1591603698988000	1591603700506000	10	6	1.5
1	Uruguay	android	1.36.7	2020-06-07 21:24:59.597 UTC	1591565099597000	1591565574073001	71	14	474.5
13	India	android	1.36.7	2020-06-08 06:02:51.517 UTC	1591596171517000	1591596325222001	22	39	153.7
9	Taiwan	ios	1.36.5	2020-06-08 11:45:45.526 UTC	1591616745526000	1591616813302001	23	10	67.8
1	India	android	1.36.7	2020-06-08 07:13:19.739 UTC	1591600399739000	1591601467677007	213	4	1067.9
1	Poland	android	1.36.6	2020-06-08 04:11:41.777 UTC	1591589501777000	1591590937822014	225	7	1436.0
2	Poland	android	1.36.6	2020-06-08 04:41:31.591011 UTC	1591591291591011	1591594984430011	476	7	3692.8
1	Ukraine	android	1.36.7	2020-06-08 17:30:40.161 UTC	1591637440161000	1591637709916006	73	1	269.8
2	United Kingdom	android	1.36.8	2020-06-07 22:03:43.087 UTC	1591567423087000	1591567944700011	66	14	521.6

Figure 6. Generated feature from daily_sessions table are in black rounded rectangle

modeling problems like unable to judge what would be sufficient, and some rarely have enough data. The supply of data for testing and training is restricted to create

useful models in some applications. It is significant to use as much data for training. If the validation set is small, it offers a noisy estimate of predictive performance.

This study has been carried out by keeping the random forest classifier as being the primacy classifier considering its capability of providing higher accuracy in research cases like this. As it has been mentioned earlier that the data set contained numerous missing values, the random forest classifier is capable enough of handling such data sets without affecting the overall performance while classification. Moreover, this classifier has the power to handle data in bulk quantity which makes it more robust.

Random forest is the supervised learning algorithm used for regression and classification. The proposed classifier tends to develop decision trees on the selected data samples and make a prediction from each tree and prefers the right solution. Even more so, studies have shown that it served as the best indicator of the feature importance. Furthermore, the random forest classifier provides an excellent feature selection indicator of the dataset. It also produces the internal unbiased count of generalization error as like the forest building process. The proposed classifier has a substantial method for balancing errors in the unbalanced data sets.

3.3.5. Feature Selection

In this section, I will select the feature from the generated features. As you can see in Figure 5 and 6, extracted ten features after analyzing the data. I choose random forest classifier for feature selection because we have a lot of data and data has some missing values and not a number values. The random forest classifier handles unbalanced data and missing values perfectly [43]. I choose features one by one and train the features on the random forest classifier and then test the accuracy of each feature. I select a threshold value that was 60% and choose the four features which have accuracy greater than or equal to 60%. The features that have accuracy higher than 60% are rew_ads_watched, ads_watched, daily_gem_state, and rew_ads_revenue. Then I make the ordered pairs of above four highest accuracy features and train again on the classifier and then a combination of three features, which gives us the accuracy 88%, and the features are rew_ads_watched, ads_watched and daily_gem_state. These features play an essential role in predicting user behavior.

3.3.6. Technical Tools

Querying a large number of datasets is a time-consuming process and expensive without the right infrastructure and hardware. To solve this issue, google big query allows super fast and SQL queries against appending mostly tables, by using the significance of google's infrastructure. I use the data from google cloud big-query for analysis. Big-Query offers logging, costly monitoring, and alerting over cloud audit logs. It also acts as the repository for records from any service or application using cloud logging. It provides a capable data repository of more high demand public sets from different organizations or industries.

Anaconda jupyter notebook is the best navigation tool to work with python. I used anaconda navigator for project work. Anaconda wants to resolve the dependency

issues in which various projects have distinct dependency versions. It is not simple to make different project dependencies need multiple versions that may interfere with each other. Jupyter will resolve the reproducibility problem in the analysis by allowing the iterative approach to highlighting code using rich text documentation with visual representations. Anaconda will attain a python environment that is 100% reproducible on the environment. In addition to this, another version of project dependencies is applicable. Jupyter is a useful presentation for analytical work where users can offer code in blocks integrates into rich text descriptions between blocks.

The programming language used in this study is python. Python is the preferred coding language used at many companies and organizations. It can be easily installed on most operating systems. It has third party libraries and earlier machine learning implementations. It is the machine learning resources developed in the collaboration. It also used as a tool for data mining and data analysis for Layman by using the application programming Interface. This programming language is compatible with significant platforms because it is used for creating or building applications. By using python interpreters, python code could run on a particular platform and support many operating systems. It is interpreted high-level programming language as it allows running code on various platforms. It helps to save the development time of developers. The code is easy to read and easily reusable. The syntax is simple that enables multiple concepts to create without any additional code. It focuses on code readability and creates custom applications. It helps to organize and update software applications.

4. EXPERIMENTAL RESULTS

We used seven days of data of racing game. We find nine features after analyzing the dataset. First of all, we train all the features separately one by one and find those features that have the highest accuracy and then make the ordered pair of all nine features and find the combination of features with higher accuracy. We used the random forest classifier for finding the best features. The parameters that we used are n_estimator that tells us the number of trees in the forest, and the value was 1000. The second parameter was random_state, and the value was none, which tells us about the controls both the randomness of the bootstrapping of the samples used when building trees. The third parameter was n_jobs, and the value of that parameter was -1, which tells us the number of jobs runs in parallel. As we are using train test split function in random forest that divides the training and testing data. So we are using 70% data for training and 30% data for testing purposes. We also used K-fold cross-validation to check our features statistically. The value of k-fold was 10 in this validation.

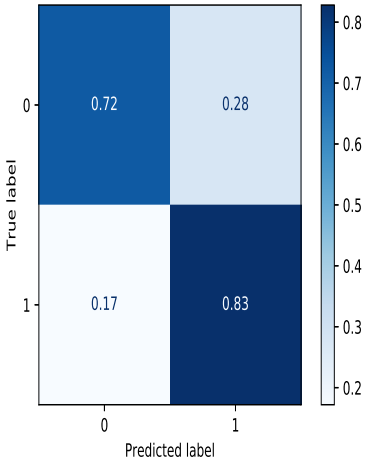

Figure 7. rew_ad_revenue

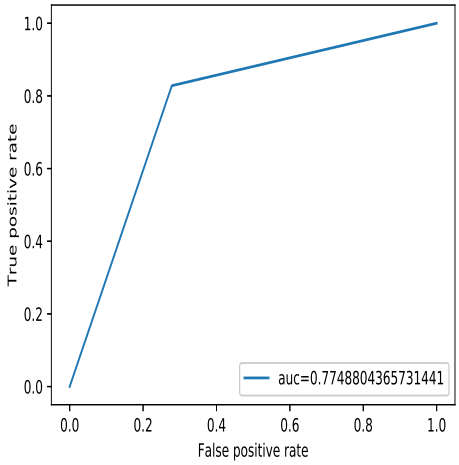

Figure 8. rew_ad_revenue

In Fig 7 and Fig 8 are the ROC curve and confusion matrix of feature 'rew_ad_revenue'. This feature tells us about the revenue that the company generated by visiting the ad by the user. We can see that in the figure 7, it clearly states that it is a useful feature because it has 72% true positive values and 83% true negative values. It can play an essential role in predicting user purchase behavior. We got 77.41% accuracy from this feature, but after the cross-validation accuracy was 74.19% where the value of K-fold was 10.

Features name	Accuracy (%)	Cross validation (K = 10)
rew_ad_revenue	77.41%	74.19%
rew_ads_watched	73%	64%
records	52.85%	54.31%
daily_gem_state	59.22%	59.54%
sess_length_seconds	52.58%	52.57%
daily_max_rank	53%	52.7%
ads_watched	85.94%	70.97
total_sessions	54%	51.77%
daily_min_rank	51.8%	51.4%
daily_coin_state	51.6%	51.2%

Table 2. Feature Set with Accuracy and K-Fold cross validation

I choose four highest accuracy features and discard the rest of the features because the other six features are not important for prediction due to less accuracy and now, will make the ordered pair of these four features and check the accuracy.

Features name	Accuracy (%)	Cross validation (K = 10)
rew_ad_revenue	77.41%	74.19%
ads_watched	85.94%	70.97
rew_ads_watched	73%	64%
daily_gem_state	59.22%	59.54%

Table 3. Feature set with highest accuracy

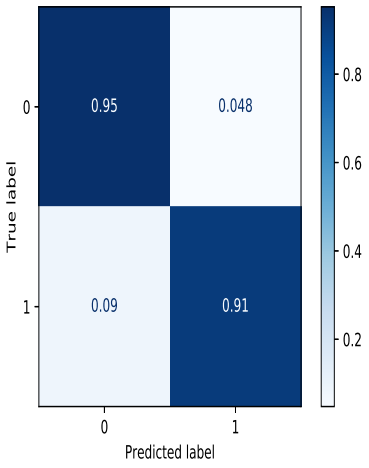

Figure 9. Confusion Matrix

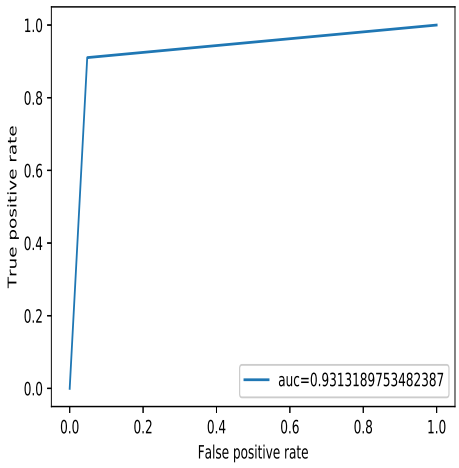

Figure 10. ROC Curve

Figure 11. rew_ad_revenue and ads_watched

In Fig 9 and Fig 10 are the ROC curve and confusion matrix of feature 'rew_ad_revenue' and 'ads_watched'. These features tells us about the generated ads revenue and ads that watched by a user in a day. In figure 9, it clearly states that the combination of these two features is essential because it has 95% true positive values and 91% true negatives values, and the value of error is very less. We got 93.14% accuracy, but after the cross-validation accuracy decreases to 83.65%.

Features name	Accuracy(%)	Cross validation (K = 10)
rew_ad_revenue ads_watched	93.14	83.65
rew_ad_revenue rew_ads_watched	74.96	72.75
rew_ad_revenue daily_gem_state	68.07	66.84
ads_watched rew_ads_watched	79.27	78.65
ads_watched daily_gem_state	83.12	70.09
rew_ads_watched daily_gem_state	79.18	78.63
rew_ad_revenue ads_watched rew_ads_watched	94.41	82.84
ads_watched rew_ads_watched daily_gem_state	90.08	86.99
rew_ad_revenue ads_watched rew_ads_watched daily_gem_state	92.35	82.97

Table 4. Feature with ordered pair and accuracy

In above all results, I train and test features one by one and then make the ordered pair of the highest accuracy features. I got a combination of three features, which gives me the highest accuracy, and they predict user behavior. As in table 4, the combination of ads watched, rew ads watched, and daily gem state gives the highest accuracy, i.e., 86.99. These three features tell us about the information about the advertisements that a user viewed and the gems that earn in one day. For graphical representation, I used the confusion matrix and ROC curve. The confusion matrix tells about the performance of a classifier on the set of text data that I have given. It is a way to visualize the performance of a classifier. The ROC curve is the graphical representation that shows the diagnostic ability of a classifier, and it discriminates against the various threshold.

4.1. Python Libraries

Python libraries are the most accessible way to perform machine learning tasks in the shortest possible time regarding the old manually coding all the algorithms, and the mathematical and statistical formula was very time-consuming. Machine learning is in the modern time, python is one of the most standard programming languages for this task, and it has substituted various languages in the industry, one of the causes is its massive collection of libraries. In the context of the implementation stage in the company, we would like to select python as the coding language that is very easy and

manageable to use on the machines. It has already installed in the system, and it is straightforward to run on the various operating system as a user need.

In the process of the machine learning algorithm, python changes everything and makes it easy for data processing in data science. The present machine learning applications prepared it perfect for data science, which contains extensive third-party libraries. The system of comprehensive analysis was to examine the sustainability of several libraries for this task, so it has been a broad drive to create python, a practical alternative in data science for more specifications like traditional tools R and MATLAB. In the time of implementation, steps on machine learning of the task are from the sci-kit-learn python library that is the most popular ML libraries for classical ML algorithms. It supports most of the supervised and unsupervised learning algorithms that can also be used for data-mining and data analysis, which makes it a great tool that is starting with ML.

Through this library could also be used for data-mining and data analysis, and the primary machine learning system that the Sci-kit-learn library can handle is classification, clustering, dimensionality reduction, regression, model selection, and preprocessing. It has inexpensive tools for data analysis and data mining readily available for layman through its application programming interface (API). It can be used for data science and machine learning tasks. ML library used in this work is MLxtend, and it implements central choice systems for machine learning and data mining for use in everyday data science tasks. It is an open-source python library that generates data structures and data analysis outfits to apply in computer science, and it also handled through Pandas.

In the context of both sci-kit-learn and the MLxtend process, pandas play a vital role in the data structures series and data frame that permit flexible storage and changes of labeled data for optimizing data analysis and modeling. There are also used several algorithms to the python libraries to scientific calculations and some statistical operations of data like NumPy, pandas, and matplotlib that are useful for data analysis from different perspectives. There are several reasons to select python for data science that is the go-to language when it brings to data science and machine learning. Python has a current-generation that most developers contribute to creating libraries for their interest and later release it to the public for their benefit.

5. DISCUSSIONS

Developing a predictive model comes with uncertain roadblocks that are meant to be resolved efficiently to set the grounds for a useful model. Likewise, there were several hindrances that were experienced during the course of this study. All of these challenges were analyzed re-actively, and techniques were employed to minimize the effects of these challenges.

The very first challenge this study had to pass was that of handling the data. It has been mentioned earlier that the data set was taken from a game that has 500 million downloads. This specific data set had instances in millions, and several data instances were having incomplete values and empty values as well. In short, data handling was the first challenge this study encountered. To streamline the data instances, the application of data preprocessing techniques was quite inevitable.

The second challenge this study faced was during the classification of the data. Before proceeding further, it is to be mentioned here that classification can only yield effective results when the classes are balanced. Any classifier whatsoever, can't be trained unless the classes are appropriately balanced, and the data is clean. If a classifier has been trained on trash data, it implies no classification task will be accomplished via that classifier because it's not appropriately trained in the first place. The classifier will give biased results, and that is not the scope of this study.

The main objective of this study, as mentioned earlier, was to predict the purchase behavior of gamers. Furthermore, breaking it down clarified the point that this study is in the pursuit of answering the question such as, whether a player would purchase the premium features or not? The answer to this question can either be affirmative or negative. From the data set this study used, around 99% of the data belonged to those gamers who would not want to purchase any of the premium features of the game, and the remaining 1% of the data set belonged to those gamers who would gladly make a purchase. Since the data set was in millions, it was decided to take a subset of this data. The data subset of 7 days was taken in such a way that fifty percent of the instances taken were positive values. And the remaining fifty percent of the negative values were selected randomly from the data, and in this way, data classes were balanced.

After balancing the data classes, the classification was done using the random forest classifier, which yielded ten features that exhibited average accuracy. It turned out to be another hindrance in the study, which was analyzed carefully, and as a result, those features were selected, having accuracy near to or above 60%. After carefully analyzing the behavior of the features, audit pairs of these features were selected, and the process of classification was again performed, which yielded three distinct features having accuracy leading up to 88%. These three features which the classifier yielded having accuracy up to 88% are: ads_watched, rew_ads_watched, and daily_gem_state.

The very first feature was ads_watched. Usually, the percentage of interest of a gamer in the game can be predicted successfully if the gamer is consistently watching the commercials. It further elaborates that the gamer wants to increase the coins or wants to upscale his inventory of gems by watching advertisements that are played at any point during the game. There are certain games that reward some coins or gems according to the number of ads that a certain gamer watches.

The second feature, which was extracted rew_ads_watched, short of rewarded advertisements watched. This feature is an extension of the feature discussed above.

There are certain games that offer reward points in the form of add-ons, coins, and gems according to the advertisements which a gamer watches without skipping. It gives an insight into the interest of the player using which we can successfully predict whether or not a gamer is going to pay a little extra for the add-ons just in case he crosses a certain level during the game.

The last feature, which was extracted as a part of this thesis, is daily_gem_state. Since there are games in which the gamer has to collect a certain amount of gems to upscale his level, it further shows that if a player is working on collecting more gems to win, he will most likely become a potential buyer of these gems once his level upscales to such a position where he sees himself as one of the pioneers of the game.

These three features, which this study extracted, can be employed in the digital marketing campaigns to make these campaigns more effective and result orienting. There can be other features as well, but for those features to come to light, another research phase will have to get started with different goals and objectives in mind. As of now, these three features have readily improved the performance of marketing campaigns and it is being hoped that a lot of budget, effort and time which once used to be spent on hit and trial basis, will be saved. The software was created in python language in jupyter notebook and integrated in company systems.

This study has primarily been done to develop an improved predictive model having the ability to predict the purchase behavior of the gamers. With a few changes and amends, the work done as a part of this study can be further reused for accomplishing the solutions of such problems, which are a part of the current digital marketing paradigm. There is a lot of room for innovative research in the field of digital marketing, gaming data analytics, and the development of predictive models, and this study only represents a subset of the whole area.

So far as the ethical challenges of this and related studies are concerned, it should be kept in mind that the sources from which the data is to be acquired, should be competent and credible enough so that any questions raised are incapable of demoralizing the overall study. Permissions to use the data in research studies should be taken from the data sources to avoid any inconveniences during or after the study has been concluded.

6. CONCLUSION

To sum up, the main objective of this study was to develop a particular model, which could benefit the digital marketing domain by plugging in the techniques which are the off-shoots of advanced technological fields like data sciences, machine learning, data mining, and data analysis. For all these fields mentioned above, the data is the only entity that holds primary importance, and in this study, the data set from a game was used.

The idea behind using the gaming data stemmed from the fact that this study aimed to develop the predictive model to be employed in marketing campaigns to predict the purchase behavior of gamers. The data which was gained for the study had a lot of incomplete instances, and the classification could only be done by simply dividing the data into YES and NO instances. After several deliberate attempts to clean and reprocess the data, it was ensured that the data no longer contains any empty instances and the data classes are equally balanced with zero chance of trash attributes or values for that matter.

A subset of the data was used in classification, training, and testing of the classifiers, which were initially, 10 in number. After analyzing, the number was reduced to the audit pairs, and the whole process of classification, training, and testing was re-employed, which gave three distinct features having comparatively improved accuracy. With these features at hand, this study concludes with the findings that these features can be used to predict the purchase behavior of gamers.

The same prediction model that we have developed can be revamped and reused by telecommunication companies in reducing the customer churn and to streamline the risk prediction. Marketing agencies can have a better insight into dealing with their newly launched products, whether physical or digital. It will pave the way for more robust marketing campaigns which can further target the exact group of buyers who can convert within a fraction of a second. It will not only lower the advertising budget for the vendor but also will help the buyer find the exact product he has been looking for all along. With the help of such prediction models, the sales and marketing funnels can be so developed that they are capable of presenting the visitor with the right popup for up-sells and down sales. The traditional propensity models are based on excel sheets, which are handled manually. Machine learning models can streamline and fine-tune such propensity models, which will lead to an agile system of marketing campaigns that can guarantee maximum profitability.

7. REFERENCES

[1] Loucks D.P. & Van Beek E. (2017) An introduction to probability, statistics, and uncertainty. In: Water Resource Systems Planning and Management, Springer, pp. 213–300.

[2] Bose I. & Mahapatra R.K. (2001) Business data mining—a machine learning perspective. Information & management 39, pp. 211–225.

[3] Koh H.C., Tan G. et al. (2011) Data mining applications in healthcare. Journal of healthcare information management 19, p. 65.

[4] Guleria P. & Sood M. (2014) Data mining in education: a review on the knowledge discovery perspective. International Journal of Data Mining & Knowledge Management Process 4, p. 47.

[5] Peña-Ayala A. (2013) Educational data mining: applications and trends, vol. 524. Springer.

[6] Clarke B., Fokoue E. & Zhang H.H. (2009) Principles and theory for data mining and machine learning. Springer Science & Business Media.

[7] Alpaydin E. (2020) Introduction to machine learning. MIT press.

[8] Rommelrath W. (2019), Predicting marketing performance with Machine Learning. https://towardsdatascience.com/predicting-marketing-performance-with-machine-learning-c8472bc7807/. [Online; accessed 22-may-2020].

[9] Trautman L.J. (2015) E-commerce, cyber, and electronic payment system risks: lessons from paypal. UC Davis Bus. LJ 16, p. 261.

[10] Ratner B. (2017) Statistical and Machine-Learning Data Mining:: Techniques for Better Predictive Modeling and Analysis of Big Data. CRC Press.

[11] Katz G., Shin E.C.R. & Song D. (2016) Explorekit: Automatic feature generation and selection. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), IEEE, pp. 979–984.

[12] Vogelsang A. & Borg M. (2019) Requirements engineering for machine learning: Perspectives from data scientists. In: 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW), IEEE, pp. 245–251.

[13] Qi Y. (2012) Random forest for bioinformatics. In: Ensemble machine learning, Springer, pp. 307–323.

[14] Oshiro T.M., Perez P.S. & Baranauskas J.A. (2012) How many trees in a random forest? In: International workshop on machine learning and data mining in pattern recognition, Springer, pp. 154–168.

[15] Yiu T. (2019), Understanding Random Forest | How the Algorithm Works and Why it Is So Effective. https://towardsdatascience.com/understanding-random-forest-58381e0602d2/. [Online; accessed 21-april-2020].

[16] Bischl B., Kerschke P., Kotthoff L., Lindauer M., Malitsky Y., Fréchette A., Hoos H., Hutter F., Leyton-Brown K., Tierney K. et al. (2016) Aslib: A benchmark library for algorithm selection. Artificial Intelligence 237, pp. 41–58.

[17] Hall M.A. (1999) Correlation-based feature selection for machine learning .

[18] Maddison C.J., Mnih A. & Teh Y.W. (2016) The concrete distribution: A continuous relaxation of discrete random variables. arXiv preprint arXiv:1611.00712 .

[19] Gonugondla S.K., Kang M. & Shanbhag N. (2018) A 42pj/decision 3.12 tops/w robust in-memory machine learning classifier with on-chip training. In: 2018 IEEE International Solid-State Circuits Conference-(ISSCC), IEEE, pp. 490–492.

[20] Andreev A. & Bogoiavlenskii I. (2017) An algorithm for building an enterprise network topology using widespread data sources. In: 2017 21st Conference of Open Innovations Association (FRUCT), IEEE, pp. 34–43.

[21] Austin P.C. & Leckie G. (2018) The effect of number of clusters and cluster size on statistical power and type i error rates when testing random effects variance components in multilevel linear and logistic regression models. Journal of Statistical Computation and Simulation 88, pp. 3151–3163.

[22] O'Madadhain J., Hutchins J. & Smyth P. (2005) Prediction and ranking algorithms for event-based network data. ACM SIGKDD explorations newsletter 7, pp. 23–30.

[23] Kh R., Deep feature synthesis is the future of machine learning. URL: https://www.smartdatacollective.com/deep-feature-synthesis-future-machine-learning.

[24] Ari B. & Güvenir H.A. (2002) Clustered linear regression. Knowledge-Based Systems 15, pp. 169–175.

[25] Chandrashekar G. & Sahin F. (2014) A survey on feature selection methods. Computers & Electrical Engineering 40, pp. 16–28.

[26] Kanter J.M. & Veeramachaneni K. (2015) Deep feature synthesis: Towards automating data science endeavors. In: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE, pp. 1–10.

[27] Bailey D.G. (2015) The advantages and limitations of high level synthesis for fpga based image processing. In: Proceedings of the 9th International Conference on Distributed Smart Cameras, pp. 134–139.

[28] Vens C. & Costa F. (2011) Random forest based feature induction. In: 2011 IEEE 11th International Conference on Data Mining, IEEE, pp. 744–753.

[29] Liaw A., Wiener M. et al. (2002) Classification and regression by randomforest. R news 2, pp. 18–22.

[30] Elssied N.O.F., Ibrahim O. & Osman A.H. (2014) A novel feature selection based on one-way anova f-test for e-mail spam classification. Research Journal of Applied Sciences, Engineering and Technology 7, pp. 625–638.

[31] Pavlidis P. (2003) Using anova for gene selection from microarray studies of the nervous system. Methods 31, pp. 282–289.

[32] Deng H. & Runger G. (2012) Feature selection via regularized trees. In: The 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1–8.

[33] Brick T.R., Koffer R.E., Gerstorf D. & Ram N. (2018) Feature selection methods for optimal design of studies for developmental inquiry. The Journals of Gerontology: Series B 73, pp. 113–123.

[34] Guyon I. & Elisseeff A. (2003) An introduction to variable and feature selection. Journal of machine learning research 3, pp. 1157–1182.

[35] Huynh-Thu V.A., Wehenkel L. & Geurts P. (2008) Exploiting tree-based variable importances to selectively identify relevant variables. Proceedings of Machine Learning Research 4, pp. 60–73.

[36] Lyon Principles of Data Mining and Knowledge Discovery.

[37] Liu H., Setiono R. et al. (1996) A probabilistic approach to feature selection-a filter solution. In: ICML, vol. 96, Citeseer, vol. 96, pp. 319–327.

[38] Hapfelmeier A. & Ulm K. (2013) A new variable selection approach using random forests. Computational Statistics & Data Analysis 60, pp. 50–69.

[39] Lessmann S. & Voß S. (2009) Feature selection in marketing applications. In: International Conference on Advanced Data Mining and Applications, Springer, pp. 200–208.

[40] Tan D.W., Sim Y.W. & Yeoh W. (2011) Applying feature selection methods to improve the predictive model of a direct marketing problem. In: International Conference on Software Engineering and Computer Systems, Springer, pp. 155–167.

[41] Kotsiantis S., Kanellopoulos D. & Pintelas P. (2006) Data preprocessing for supervised learning. International Journal of Computer Science 1, pp. 111–117.

[42] Morales-Menendez R., J.A y., Cantu-Ortiz F. & Cavazos J. (2005) Model selection in data mining: A statistical approach.

[43] Kho J. (2018), Why Random Forest is My Favorite Machine Learning Model. https://https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706/. [Online; accessed 10-june-2020].